

TESTOWANIE WIELOKROTNE W BADANIACH EKONOMICZNYCH

SABINA DENKOWSKA

Katedra Statystyki
Uniwersytet Ekonomiczny w Krakowie
PL 31-150 Kraków, ul. Rakowicka 27
e-mail: sabina.denkowska@ae.krakow.pl

Praca przedstawiona na posiedzeniu Komisji Nauk Ekonomicznych i Komisji Statystyczno-Demograficznej Oddziału PAN w Krakowie 22 maja 2007 roku przez autorkę.

ABSTRACT

Sabina Denkowska. *Multiple testing in the Economic Research*. Folia Oeconomica Cracoviensia 2007, 48, 119–135.

In the paper multiple testing procedures based on *ordered p-values* are applied to control type I errors rates for family of inferences in different statistical problems. Multiple testing procedures based on *ordered p-values* may be found an interesting tool for simultaneous testing of more than one hypothesis at a time. Discussed methods are applicable to a broad spectrum of statistical problems since their requirements for statistical assumption are considerably less restricted than in case of classical procedures (only dependency among test statistics should be controlled). The analysis is put down to a collection of p-values or adjusted p-values. Depending on approach to the control of Type I error for the family of inferences these methods may be categorized into two major groups: FWE (Family-Wise Error Rate) and FDR (False Discovery Rate) procedures.

The procedures of multiple testing are applied to typical situations in economics research: to separate homogenous groups of means, to test the significance of correlation coefficients in the correlation matrix and to infer about significance of regression parameters in linear regression model.

KEY WORDS — SŁOWA KLUCZOWE

multiple testing, multiple comparison procedures MCP, simultaneous testing, multiple testing procedures based on ordered p-values, adjusted p-values, FWE (Family-Wise Error Rate), FDR (False Discovery Rate)

testowanie wielokrotne, procedury porównań wielokrotnych, testowanie jednoczesne (równoległe), procedury testowań wielokrotnych oparte na prawdopodobieństwach testowych, skorygowane prawdopodobieństwa testowe, FWE, FDR

WSTĘP

Kontrolą efektu równoczesnego testowania wielu hipotez zajmuje się teoria wnioskowań wielokrotnych. Dynamiczny rozwój tej dziedziny wiedzy wynika z intensywnego stosowania metod statystycznych w różnorodnych dziedzinach nauki: medycynie, farmacji, genetyce, fizyce, astronomii itp. Również w badaniach ekonomicznych testowanie wielokrotne dotyczy bardzo wielu sytuacji badawczych. W przeciwieństwie do wnioskowania sekwencyjnego, które zazwyczaj występuje *explicite*, równoległe testowanie wielokrotne jest często niedostrzegane, a czasem po prostu świadomie lekceważone.

Zazwyczaj wnioski z badań empirycznych wysnuwane są po zweryfikowaniu wielu hipotez statystycznych, każdej na poziomie istotności α . Tym samym ignorowany jest fakt, że im większa liczba rozpatrywanych hipotez, tym jest większe prawdopodobieństwo wykrycia pozornie istotnych statystycznie związków, które w rzeczywistości nie istnieją. Dla zilustrowania problemu rozważmy k niezależnych testów istotności, których hipotezy zerowe są prawdziwe. Testowania przeprowadzamy na poziomie istotności α . Prawdopodobieństwo tego, że nie odrzucimy żadnej hipotezy zerowej wynosi $(1 - \alpha)^k$. Jeżeli przykładowo, testujemy $k = 20$ niezależnych, prawdziwych hipotez zerowych, każdą na poziomie istotności 0,05 to prawdopodobieństwo odrzucenia przynajmniej jednej z nich wynosi $1 - 0,95^{20} = 0,64$, czyli okazuje się, że bardziej prawdopodobne jest uzyskanie przynajmniej jednego istotnego statystycznie wyniku, niż nie odrzucenie żadnej z hipotez zerowych. Ponieważ w badaniach praktycznych rzadko mamy do czynienia z niezależnymi testami, więc powyższe obliczenia należy traktować jako ostrzeżenie mówiące o tym, że im więcej testowań przeprowadzamy, tym więcej występuje błędnych odrzuceń hipotez zerowych. Należy zatem starać się ograniczać w badaniach liczbę przeprowadzanych testowań oraz stosować właściwą metodologię, umożliwiającą kontrolowanie opisanego powyżej efektu wnioskowania wielokrotnego.

Klasyczne procedury wnioskowań wielokrotnych MCP (ang. *Multiple Comparison Procedure*), od lat obecne w literaturze naukowej, mają liczne grono zarówno zwolenników, jak i przeciwników. Już w najbardziej typowej sytuacji wnioskowania wielokrotnego przy wydzielaniu jednorodnych podgrup wartości przeciętnych, gdy ANOVA spowoduje odrzucenie hipotezy zerowej, badacz napotyka problem wyboru najwłaściwszej procedury z bogatego zbioru procedur *post-hoc*, zwłaszcza że różne procedury dają różne podziały zbioru wartości przeciętnych (Denkowska 1999). Kolejną wadą procedur *post-hoc* są otrzymywane wyniki, częstokroć trudne do zinterpretowania na gruncie praktyki (np. nierozłączne jednorodne podzbiory wartości przeciętnych). Rozwiązania klasyczne zalecane w literaturze przedmiotu mają ponadto rygorystyczne wymagania modelowe.

Wybór procedury MCP zależy m.in. od zbioru elementów, które zamierzamy testować. Do typowych sytuacji badawczych należą: porównywanie parami

wartości przeciętnych, porównywanie parami wartości przeciętnych z wartością kontrolną lub „najlepszą” oraz testowanie istotności kontrastów, czyli kombinacji liniowych wartości przeciętnych, dla których suma współczynników wynosi zero. Innym bardzo istotnym czynnikiem decydującym o wyborze procedury są założenia modelu statystycznego. Wybór właściwej procedury nie jest zadaniem prostym. Wielość zaawansowanych i złożonych procedur, różnorodność uzyskiwanych wyników w zależności od wybranej procedury oraz trudności interpretacyjne mogą zniechęcać praktyków do stosowania procedur wnioskowań wielokrotnych. A przecież testowanie wielokrotne to nie tylko porównywanie wartości przeciętnych wymienione wcześniej, czy testowanie istotności kontrastów. Z testowaniem wielu hipotez mamy do czynienia np. przy doborze zmiennych objaśniających do modelu regresji podczas testowania istotności współczynników korelacji w macierzy korelacji lub przy testowaniu istotności ocen parametrów strukturalnych w modelu regresji wielorakiej.

Alternatywę dla procedur klasycznych stanowią procedury testowań wielokrotnych, oparte na prawdopodobieństwach testowych, w których proces testowania opiera się przede wszystkim na analizie indywidualnych prawdopodobieństw testowych. Procedury testowań wielokrotnych oparte na uporządkowanych prawdopodobieństwach testowych rekomendowane są w licznych artykułach naukowych, zarówno z genetyki, psychologii, jak i farmacji, czy meteorologii. Zaletą tych procedur jest fakt, iż zakres ich zastosowań może być bardzo szeroki, a ich wymagania co do założeń modelu statystycznego są nieduże. Istotną kwestią jest jedynie typ zależności pomiędzy statystykami testowymi, gdyż w przypadku niezależnych statystyk testowych lub pewnych typów zależności należy stosować procedury o większej mocy.

Testowanie wielokrotne w badaniach ekonomicznych jest faktem. W dobie wielkiej popularności metody *data mining*, kontrola efektu testowania wielokrotnego wydaje się po prostu niezbędna. Nasuwa się zatem pytanie: czy kontrolę efektu wnioskowania wielokrotnego w badaniach ekonomicznych można powierzyć procedurom testowań wielokrotnych opartym na prawdopodobieństwach testowych.

W artykule rozważane jest stosowanie procedur testowań wielokrotnych opartych na uporządkowanych prawdopodobieństwach testowych w typowych sytuacjach badawczych występujących w ekonomii:

- do porównywania parami wartości przeciętnych w celu wydzielania jednorodnych podgrup wartości przeciętnych w modelu jednoczynnikowej analizy wariancji;
- do testowania istotności współczynników korelacji w macierzy korelacji;
- do testowania istotności parametrów strukturalnych w modelu regresji wielorakiej.

2. WPROWADZENIE DO TEORII TESTOWAŃ WIELOKROTNYCH

Właściwa kontrola efektu wielokrotności testowania jest zagadnieniem trudnym i kontrowersyjnym. Pierwszą, a zarazem kluczową decyzją jest wybór zbioru wnioskowań, który będzie tworzyć rodzinę wnioskowań. Hochberg i Tamhane (1987) zalecają, żeby w przypadku, gdy testowane hipotezy nie są ze sobą powiązane ani zawartością, ani późniejszym wykorzystaniem, traktować je oddzielnie, a nie łącznie. W przeciwnym wypadku, istotnym jest branie pod uwagę łącznego pomiaru błędów. Gdy wniosek końcowy wysnuwany jest na podstawie przeprowadzonych testów analizowanych łącznie i jego trafność zależy od łącznego pomiaru błędów dla danego zbioru wnioskowań, wtedy taki zbiór wnioskowań powinien być rozpatrywany łącznie jako rodzina.

W celu zaprezentowania najważniejszych miar błędu I rodzaju dla rodziny wnioskowań rozważmy problem jednoczesnego testowania k hipotez, wśród których k_0 hipotez jest prawdziwych. Niech R oznacza liczbę hipotez zerowych odrzuconych na korzyść odpowiednich hipotez alternatywnych, zaś V — liczbę prawdziwych hipotez zerowych, które odrzucamy. R , V są to zmienne losowe. Po przeprowadzeniu testowania znana jest tylko liczba hipotez, które odrzucamy — R , a tym samym liczba hipotez zerowych, dla których nie mamy podstaw do odrzucenia: $k - R$. Wartość zmiennej losowej V nie jest obserwowana.

W literaturze tematu najczęściej wyróżniane są dwie miary błędu I rodzaju dla rodziny wnioskowań FWE (ang. *Family-Wise Error Rate*) (Hochberg i Tamhane, 1987; Miller, 1981):

$$\text{FWE} = P(V \geq 1),$$

FDR (ang. *False Discovery Rate*):

$$\text{FDR} = \begin{cases} E\left(\frac{V}{R}\right) & \text{gdy } R > 0 \\ 0 & \text{gdy } R = 0. \end{cases}$$

Kontrola FWE dla rodziny wnioskowań oznacza, iż prawdopodobieństwo odrzucenia przynajmniej jednej prawdziwej hipotezy zerowej jest nie większe od ustalonego z góry α . Nie we wszystkich badaniach stosowanie procedur kontrolujących FWE daje dobre rezultaty. W przypadku licznych rodzin wnioskowań, procedury FWE osłabiają moc indywidualnych wnioskowań, w wyniku czego otrzymujemy zbyt mało odrzuceń hipotez zerowych. W 1995 roku Benjamini i Hochberg zaproponowali nową miarę błędu I rodzaju dla rodziny wnioskowań FDR (ang. *False Discovery Rate*), której kontrola oznacza, iż wartość oczekiwana frakcji błędnych odrzuceń wśród wszystkich odrzuceń hipotez zerowych jest kontrolowana na ustalonym z góry poziomie.

Dla zilustrowania różnicy pomiędzy FWE i FDR rozpatrzmy sytuację, gdy rodzina wnioskowań obejmuje 1000 hipotez zerowych i odpowiadających im hipotez alternatywnych. Porównajmy sytuację odrzucenia 100 hipotez zerowych, z czego jedna jest prawdziwa, z sytuacją odrzucenia w tym zbiorze dwóch hipotez, z czego jedna jest prawdziwa. Z punktu widzenia FWE obie te sytuacje są tak samo niekorzystne, bo odrzucona została jedna prawdziwa hipoteza zerowa. Można jednak spojrzeć na ten wynik w ten sposób, iż tylko 1% odrzuceni było błędnych w pierwszej sytuacji, a aż 50% w drugiej.

3. PROCEDURY TESTOWAŃ WIELOKROTNYCH OPARTE NA PRAWDOPODOBIENSTWACH TESTOWYCH

Procedury testowań wielokrotnych oparte na prawdopodobieństwach testowych wydają się być interesującą alternatywą dla klasycznych procedur MCP. Zakres zastosowań tych procedur jest bardzo szeroki. Procedury te mogą być stosowane w przypadku skończonych rodzin hipotez minimalnych, a proces testowania przy ich wykorzystaniu opiera się przede wszystkim na analizie indywidualnych prawdopodobieństw testowych. Istotną zaletą tych procedur są nieduże wymagania co do założeń modelu statystycznego.

Z procedurami testowań wielokrotnych ściśle związane jest pojęcie skorygowanych prawdopodobieństw testowych (ang. *adjusted p-values*). Analogicznie do definicji zwykłych (nieskorygowanych) prawdopodobieństw testowych p , skorygowane prawdopodobieństwo \tilde{p}_i dla hipotezy $H_{0,i}$ vs. $H_{A,i}$, równe jest najmniejszej wartości FWE, dla której $H_{0,i}$ może zostać odrzucona, gdy cała rodzina hipotez jest rozpatrywana. Przy czym zakładamy, że rozpatrujemy rodzinę 'k' minimalnych hipotez zerowych $H_{0,1}, H_{0,2}, \dots, H_{0,k}$ z odpowiadającymi im prawdopodobieństwami testowymi p_1, p_2, \dots, p_k . Mając wyznaczone¹ skorygowane prawdopodobieństwa testowe \tilde{p}_i ($i = 1, \dots, k$) dla każdego testu $H_{0,i}$ vs. $H_{A,i}$, decyzja o odrzuceniu hipotezy $H_{0,i}$ na poziomie FWE równym α podejmowana jest, gdy $\tilde{p}_i \leq \alpha$. Skorygowane prawdopodobieństwa testowe są definiowane analogicznie dla FDR.

Ze względu na kontrolę błędu I rodzaju dla rodziny wnioskowań, procedury testowań wielokrotnych można podzielić na:

- procedury kontrolujące FWE;
- procedury kontrolujące FDR.

Do uniwersalnych procedur testowań wielokrotnych kontrolujących FWE zaliczamy jednoetapową procedurę Bonferroniego oraz jej wieloetapową mo-

¹ Sposoby wyznaczania skorygowanych prawdopodobieństw testowych dla poszczególnych procedur testowań wielokrotnych przedstawione zostaną w dalszej części artykułu.

dyfikację — procedurę Holma. Procedura Bonferroniego jest najprostszą procedurą testowań wielokrotnych, a jej algorytm można przedstawić następująco:

$$\text{odrzucaamy } H_{0,i} \text{ wtedy gdy } p_i \leq \frac{\alpha}{k}.$$

Metoda Bonferroniego jest bardzo konserwatywna, czyli jest metodą o malej mocy. Skorygowane prawdopodobieństwa testowe dla metody Bonferroniego wyznaczane są ze wzoru:

$$\tilde{p}_j = \min(kp_j; 1) \quad \text{dla } j = 1, \dots, k. \quad (1)$$

Procedura Holma jest mniej konserwatywna niż metoda Bonferroniego, bowiem każda hipoteza odrzucona przez metodę Bonferroniego jest odrzucona również przez metodę Holma, natomiast hipotezy odrzucone przez metodę Holma mogą nie zostać odrzucone przez metodę Bonferroniego. Dla uporządkowanych prawdopodobieństw testowych $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(k)}$ skorygowane prawdopodobieństwa testowe dla metody Holma wyznaczane są ze wzorów:

$$\begin{aligned} \tilde{p}_{(1)} &= \min(1; kp_{(1)}) \quad \text{oraz} \\ \tilde{p}_{(j)} &= \min\left(1; \max(\tilde{p}_{(j-1)}, (k-j+1)p_{(j)})\right) \\ &\quad \text{dla } j = 2, \dots, k. \end{aligned} \quad (2)$$

Procedurę Bonferroniego można zmodyfikować korzystając z nierówności Šidáka w przypadku, gdy rozpatrywane statystyki testowe tworzą wielowymiarowy rozkład normalny lub rozkład *t*-Studenta o niezależnych składowych (Hochberg i Tamhane, 1987; Shaffer, 1995), a rozważane hipotezy alternatywne mają dwustronne zbiory krytyczne. Modyfikacja metody Bonferroniego polega na zastąpieniu $\frac{\alpha}{k}$ przez $1 - (1 - \alpha)^{\frac{1}{k}}$. Prawdopodobieństwa skorygowane dla procedury Bonferroniego-Šidáka liczone są ze wzoru:

$$\tilde{p}_j = \min\left(1 - (1 - p_j)^k; 1\right) \quad \text{dla } j = 1, \dots, k. \quad (3)$$

Jak wykazali Holland i Copenhaver (1987) procedura Bonferroniego-Šidáka kontroluje FWE również w przypadku, gdy statystyki testowe mają dodatnią zależność orthantową (Denuit i Scaillet, 2004).

Procedura Holma-Šidáka jest modyfikacją metody Holma opartą na nierówności Šidáka. Skorygowane prawdopodobieństwa testowe dla tej metody wyznaczane są ze wzorów:

$$\begin{aligned}\tilde{p}_{(1)} &= \min\left(1; 1 - (1 - p_{(1)})^k\right) \text{ oraz} \\ \tilde{p}_{(j)} &= \min\left(1; \max\left(\tilde{p}_{(j-1)}, 1 - (1 - p_{(j)})^{k-j+1}\right)\right) \\ &\text{dla } j = 2, \dots, k.\end{aligned}\quad (4)$$

Procedura Shaffer to modyfikacja procedury Holma dla hipotez logicznie powiązanych. Z hipotezami logicznie powiązanymi mamy do czynienia np. w przypadku porównywania parami wartości przeciętnych. Prawdopodobieństwa skorygowane dla tej metody wyznaczamy ze wzorów:

$$\tilde{p}_{(1)} = \min(1; kp_{(1)}) \text{ oraz } \tilde{p}_{(j)} = \min\left(1; \max(\tilde{p}_{(j-1)}, t_j p_{(j)})\right) \text{ dla } j = 2, \dots, k, \quad (5)$$

gdzie t_j — jest to maksymalna liczba hipotez zerowych, dla których możemy stwierdzić brak podstaw do odrzucenia, gdy $(j - 1)$ — hipotez jest już odrzuconych.

Kontrolę FDR dla niezależnych statystyk testowych zapewnia procedura LSU (*Linear Step-Up*), zaproponowana przez Hochberga i Benjaminiego (1995). Skorygowane prawdopodobieństwa testowe dla procedury Hochberga-Benjaminiego otrzymujemy ze wzorów:

$$\tilde{p}_{(k)} = p_{(k)}; \quad \tilde{p}_{(k-j)} = \min\left(\tilde{p}_{(k-j+1)}, \frac{k}{k-j} p_{(k-j)}\right) \text{ dla } j = 1, \dots, k - 1. \quad (6)$$

Benjamini i Yekutieli (2001) wykazali, że procedura LSU zapewnia kontrolę FDR również w przypadku statystyk testowych o zależności dodatnio regresyjnej.

Uniwersalną procedurą kontrolującą FDR jest procedura Yekutieli-Benjaminiego². Jest ona modyfikacją procedury LSU (*Linear Step-Up*) Hochberga-Benjaminiego zapewniającą kontrolę FDR bez względu na typ zależności pomiędzy statystykami testowymi. Niestety procedura ta jest procedurą konserwatywną w porównaniu z procedurą LSU. Skorygowane prawdopodobieństwa testowe dla metody Yekutieli-Benjaminiego wyznaczamy ze wzorów:

$$\begin{aligned}\tilde{p}_{(k)} &= \min\left(1; p_{(k)} \sum_{i=1}^k \frac{1}{i}\right) \text{ oraz} \\ \tilde{p}_{(k-j)} &= \min\left(\tilde{p}_{(k-j+1)}, p_{(k-j)} \frac{k}{k-j} \sum_{i=1}^k \frac{1}{i}\right) \\ &\text{dla } j = 1, \dots, k - 1.\end{aligned}\quad (7)$$

² Patrz Benjamini i Yekutieli (2001).

4. WYDZIELANIE JEDNORODNYCH PODGRUP WARTOŚCI PRZECIĘTNYCH

Przydatność procedur testowań wielokrotnych opartych na prawdopodobieństwach testowych do wydzielania jednorodnych podgrup wartości przeciętnych badano poprzez konfrontację tych procedur z procedurami klasycznymi w modelu jednoczynnikowej zrównoważonej analizy wariancji.

Klasyczne procedury *post-hoc*, procedury oparte na analizie skupień oraz wybrane, na podstawie teoretycznych rozważań, procedury testowań wielokrotnych oparte na prawdopodobieństwach testowych, zastosowano w przykładach empirycznych dotyczących wyodrębniania grup gospodarstw domowych o jednorodnym poziomie przeciętnych wydatków na osobę na wybrane artykuły. Badania empiryczne zaprezentowane przez autorkę (Denkowska, 2005) pokazały typowe problemy napotymane przy stosowaniu klasycznych procedur *post-hoc*. Procedury porównań wielokrotnych od lat dostępne w pakietach statystycznych nadal budzą sporo kontrowersji. Różnorodność otrzymywanych wyników, nierozłączność jednorodnych podzbiorów wartości oczekiwanych, to tylko niektóre z czynników zniechęcających badaczy do sięgania po te procedury. Procedury oparte na uporządkowanych prawdopodobieństwach testowych wypadły w badaniach porównywalnie, a czasami nawet lepiej niż rozwiązania klasyczne (Denkowska, 2005). Na szczególne wyróżnienie zasługiwała procedura Holma, która niejednokrotnie dawała więcej odrzuceń hipotez zerowych, niż zalecana w literaturze do tego typu badań procedura Tukeya. Biorąc pod uwagę rezultaty badań empirycznych oraz zalety procedur testowań wielokrotnych opartych na prawdopodobieństwach testowych, takie jak szeroki zakres ich zastosowań oraz prostotę ich stosowania, kontynuowano badania nad efektywnością i wykorzystaniem tych procedur do bardziej zaawansowanych problemów testowań wielokrotnych.

W celu porównania efektywności klasycznych procedur *post-hoc* z procedurami testowań wielokrotnych opartymi na prawdopodobieństwach testowych w modelu statystycznym jednoczynnikowej zrównoważonej analizy wariancji, autorka artykułu (Denkowska, 2006b) przeprowadziła rozbudowane badania Monte Carlo. Badania te polegały na generowaniu prób z rozkładów normalnych o znanych parametrach, a następnie badaniu, czy decyzja podjęta na podstawie tych prób jest poprawna, czy nie. Próby takie były generowane wielokrotnie, a prawdopodobieństwo poprawnej decyzji było szacowane jako częstość występowania dobrych decyzji.

Procedury poddane badaniom symulacyjnym można podzielić na trzy grupy. Jedną grupę procedur stanowiły najpopularniejsze procedury klasyczne dostępne w pakiecie STATISTICA: procedura NIR, procedura Tukeya, procedura Scheffégo, procedura Duncana oraz procedura Newmana-Keulsa. Pozostałe procedury to wybrane procedury testowań wielokrotnych oparte na uporządkowanych prawdopodobieństwach testowych, które można zastosować w przypadku modelu jednoczynnikowej zrównoważonej analizy wariancji do wydzielania

jednorodnych podgrup. Grupę drugą — grupę procedur kontrolujących prawdopodobieństwo popełnienia przynajmniej jednego błędu I rodzaju na poziomie α (FWE), tworzyły: procedura Bonferroniego, procedura Bonferroniego-Šidaka, procedura Holma, procedura Holma-Šidaka oraz procedura Shaffer dla hipotez logicznie powiązanych. W skład grupy trzeciej wchodziły procedury kontrolujące FDR, czyli wartość oczekiwaną frakcji błędnych odrzuceń wśród wszystkich odrzuceń. Badaniom poddano: procedurę LSU Hochberga-Benjaminiego³ oraz procedurę Yekutieli-Benjaminiego. Tak więc badaniami Monte Carlo objęto 12 procedur porównań wielokrotnych. Program symulacyjny porównujący 12 procedur MCP napisano w języku STATISTICA Visual Basic (SVB). W programie wykorzystano gotowe, oprogramowane w STATISTICE, popularne procedury klasyczne z pierwszej grupy procedur. Procedury z dwóch pozostałych grup zostały oprogramowane w oparciu o wyznaczane w STATISTICE prawdopodobieństwa testowe dla par wartości przeciętnych.

Zarówno przeprowadzone badania symulacyjne, jak i badania empiryczne pokazały, że procedury testowań wielokrotnych oparte na prawdopodobieństwach testowych mogą stanowić poważną konkurencję dla procedur klasycznych. W przypadku procedur kontrolujących FWE w rozważanym w pracy modelu statystycznym zrównoważonej analizy wariancji, efektywność procedur opartych na prawdopodobieństwach testowych była porównywalna z efektywnością procedur klasycznych. Spośród procedur klasycznych wysoką efektywnością charakteryzowały się procedury Newman-Keulsa i Duncana, jednak stosowanie tych procedur nie jest zalecane w praktyce badawczej, gdyż nie kontrolują one FWE na zadanym poziomie. Szczególnie zalecana w literaturze przedmiotu do porównań parami w rozważanym modelu statystycznym jest procedura Tukeya. W prowadzonych badaniach symulacyjnych dobrze wypadły wieloetapowe procedury testowań wielokrotnych, osiągając niejednokrotnie efektywność wyższą od efektywności klasycznej procedury Tukeya. Badania Monte Carlo nie stanowią formalnego dowodu, jednak biorąc dodatkowo pod uwagę fakt, iż procedury takie jak np. procedura Holma, czy jej modyfikacja zaproponowana przez Shaffer mogą być stosowane do porównywania parami wartości przeciętnych również w sytuacjach, gdy założenia modelowe analizy wariancji nie są spełnione, z pewnością warto promować tę grupę metod porównań wielokrotnych.

Podsumowując wyniki empiryczne i symulacyjne dotyczące procedur testowań wielokrotnych opartych na prawdopodobieństwach testowych, można stwierdzić, że spośród metod kontrolujących FWE na szczególne wyróżnienie zasługuje wieloetapowa metoda Holma. Zarówno modyfikacja zaproponowana przez Shaffer, polegająca na uwzględnieniu logicznych powiązań pomiędzy

³ Procedura Hochberga-Benjaminiego, na podstawie badań symulacyjnych przeprowadzonych przez Y. Benjaminiego, Y. Hochberga oraz Y. Klinga, zapewnia kontrolę FDR w przypadku rozpatrywanych w eksperymencie porównań parami. Więcej na ten temat u Benjaminiego i Yekutieli, 2001.

hipotezami, jak również modyfikacja wykorzystująca nierówność Šidáka, nie wpłynęły znacząco na poprawę efektywności procedury Holma. Teoretyczna analiza prawdopodobieństw skorygowanych w przypadku procedury Holma-Šidáka nie zapowiadała rewolucyjnych zmian i fakt ten potwierdziły badania symulacyjne. Zastosowanie nierówności Šidáka skomplikowało natomiast prostotę obliczeniową metody Holma, a uzyskana tą drogą korekta wartości skorygowanych okazała się nieznaczna, przy czym modyfikacja ta zapewnia kontrolę FWE tylko w przypadku niezależnych lub charakteryzujących się dodatnią zależnością orthantową statystyk testowych. Właściwie niezauważalna jest poprawa efektywności metody Holma w wyniku zastosowania modyfikacji Shaffer. Wydaje się jednak, że korzyść z zastosowania tej poprawki dla hipotez logicznie powiązanych będzie widoczniejsza przy rozpatrywaniu większej liczby parametrów. W przypadku rozpatrywania znacznej liczby wartości przeciętnych, gdy w rozpatrywanych badaniach wystarczy kontrola wartości oczekiwanej frakcji błędnych odrzuceń wśród wszystkich odrzuceń na z góry zadanim poziomie, warto polecić procedurę LSU Hochberga-Benjaminiego. Natomiast uniwersalna procedura Yekutieli-Benjaminiego okazała się procedurą bardzo konserwatywną i to nie tylko w stosunku do procedury LSU.

5. TESTOWANIE WIELOKROTNE W MACIERZY KORELACJI

W wielu praktycznych zastosowaniach statystyki i ekonometrii zachodzi konieczność wnioskowania o istotności współczynników korelacji liniowej w macierzy korelacji. Podzbiorów wzajemnie skorelowanych zmiennych poszukujemy testując wielokrotnie w macierzy korelacji, m.in. przy wyborze zmiennych objaśniających do modelu ekonometrycznego, zmiennych definiujących przestrzeń klasyfikacji w analizie skupień, zmiennych tworzących wskaźniki agregatywne, zmiennych diagnostycznych w analizie dyskryminacyjnej. We wszystkich wymienionych przypadkach zazwyczaj ignorowany jest fakt testowania wielokrotnego. Dostępne w literaturze klasyczne testy globalne dotyczące współczynników korelacji liniowej (Domański, 1990) pozostawiają badacza w kłopotliwej sytuacji, gdy zostanie odrzucona hipoteza zerowa. Wówczas, aby wysnuć bardziej szczegółowe wnioski i wskazać istotne statystycznie współczynniki korelacji, najczęściej stosowanym rozwiązaniem jest testowanie istotności współczynników korelacji, każdego na poziomie istotności α . Brak klasycznych rozwiązań kontrolujących efekt testowania wielokrotnego w przypadku badania istotności współczynników korelacji w macierzy korelacji powoduje, że przydatność procedur testowań wielokrotnych opartych na prawdopodobieństwach testowych jest bezsporna.

Testowanie z wykorzystaniem procedur opartych na prawdopodobieństwach testowych należy rozpocząć od określenia rodziny wnioskowań. Bardzo ważne jest,

by na wstępie badań starać się poprzez analizę merytoryczną oraz analizę formalno-statystyczną maksymalnie zredukować liczbę wnioskowań, czyli w przypadku testowania wielokrotnego w macierzy korelacji — zredukować liczbę zmiennych. Po przeprowadzeniu analizy merytorycznej i formalno-statystycznej zostaje ustalona liczba rozważanych zmiennych, a tym samym znana jest rodzina wnioskowań. Kolejny etap polega na wyznaczeniu prawdopodobieństw testowych dla całej rozważanej rodziny wnioskowań, a następnie na zastosowaniu „najlepszej” z procedur testowań wielokrotnych opartych na uporządkowanych prawdopodobieństwach testowych. Wybór „najlepszej” procedury rozumiany jest jako wskazanie procedury o największej mocy, w zależności od wyboru miary kontroli błędu I rodzaju dla całej rodziny wnioskowań. W przypadku kontroli FWE w macierzy korelacji, wybór zawęży się do dwóch metod: Bonferroniego oraz metody „sekwencyjnego odrzucania” Holma, które można polecić bez względu na typ zależności pomiędzy statystykami testowymi. Metodą większej mocy spośród rozważanych, jest metoda Holma, bowiem w wyniku jej zastosowania liczba wykrytych istotnych współczynników korelacji będzie nie mniejsza od liczby wykrytych istotnych współczynników korelacji przez metodę Bonferroniego, przy zapewnieniu kontroli prawdopodobieństwa błędnego stwierdzenia jednej lub więcej istotnych współzależności pomiędzy zmiennymi na założonym poziomie α dla całej rodziny wnioskowań. W przypadku, gdy zalecana redukcja zmiennych nie spowoduje znacznego zmniejszenia liczby cech i rodzina wnioskowań jest bardzo liczna, warto rozważyć kontrolę FDR przy pomocy metody Yekutieli-Benjaminiego. Dla dużych macierzy korelacji kontrola FWE nie jest rozwiązaniem satysfakcjonującym badaczy. Przy licznej rodzinie wnioskowań indywidualne poziomy istotności, w przypadku procedur FWE, są tak małe, iż zbyt rzadko dochodzi do odrzucenia hipotez zerowych i stwierdzania istotnych współzależności pomiędzy zmiennymi. Tak więc w sytuacji dużych macierzy korelacji, jeśli dopuszczamy $\alpha 100\%$ — błędnych odrzuceń wśród wszystkich odrzuceń hipotez zerowych, warto wykorzystać procedurę Yekutieli-Benjaminiego, kontrolującą FDR na poziomie α . W praktyce oznacza to, że akceptujemy niewielki procent współczynników błędnie uznanych za istotne, wśród wszystkich uznanych za istotne współczynników korelacji.

W przykładzie empirycznym, zaprezentowanym u autorki (Denkowska, 2006a), wyznaczano cechy diagnostyczne do oceny przestrzennego zróżnicowania sytuacji mieszkaniowej w Polsce w 2002 roku. Przedstawiono w nim dwa podejścia: jedno polegające na tradycyjnym wnioskowaniu w macierzy korelacji, w którym ignorowany jest efekt wielokrotności testowania oraz drugie podejście, w którym efekt ten kontrolowano za pomocą wybranych metod testowań wielokrotnych opartych na prawdopodobieństwach testowych. Do kontroli efektu wielokrotności testowania zastosowano wybrane uniwersalne procedury kontrolujące FWE i FDR. Przykład empiryczny pokazał, że lekceważenie efektu testowania wielokrotnego powoduje wykrywanie zbyt wielu istotnych współ-

czynników korelacji, co w rezultacie skutkuje zbytnią redukcją zmiennych diagnostycznych. Tym samym, w finalnym zbiorze zmiennych może zabraknąć ważnych zmiennych diagnostycznych.

Dobór zmiennych objaśniających do modeli ekonometrycznych jest jednym z najistotniejszych zagadnień współczesnej ekonometrii. Od lat bardzo popularną metodą doboru zmiennych objaśniających do liniowego modelu regresji jest metoda oparta na badaniu istotności współczynników korelacji w macierzy korelacji (Bartosiewicz, 1980; Grabiński i wsp., 1982). Metoda ta jest spotykana w literaturze pod nazwą metody analizy grafów (Goryl i wsp., 2003) i znajduje zastosowanie przede wszystkim w przypadku, gdy rozpatrywana jest niezbyt liczna rodzina potencjalnych zmiennych objaśniających. W metodzie tej ignorowany jest fakt testowania wielokrotnego w macierzy korelacji i wnioski wysnuwane są po zweryfikowaniu wielu hipotez statystycznych, każdej na poziomie istotności α . Lekceważenie efektu testowania wielokrotnego powoduje zbytnią redukcję zmiennych objaśniających.

Autorka (Denkowska, 2007a) zaproponowała modyfikacje metody analizy grafów uwzględniające efekt wielokrotności testowania w macierzy korelacji. Modyfikacje te polegają na wykorzystaniu odpowiednich procedur testowań wielokrotnych kontrolujących FWE lub FDR do kontroli efektu testowania wielokrotnego istotności współczynników korelacji liniowej w macierzy korelacji. W przypadku, gdy istotne jest, by prawdopodobieństwo błędnego stwierdzenia jednej lub więcej istotnych statystycznie zależności było kontrolowane na poziomie α , najlepszym rozwiązaniem jest zastosowanie modyfikacji opartej na metodzie Holma. Kontrola FWE na poziomie α oznacza, że uzyskane istotne wyniki możemy z dużym zaufaniem przyjąć, gdyż prawdopodobieństwo błędnego alarmu jest nie większe od α . Natomiast w sytuacji licznej rodziny wnioskowań warto rozważyć, czy nie wystarczy kontrola (FDR) wartości oczekiwanej frakcji błędnych odrzuceń wśród wszystkich odrzuceń na z góry ustalonym poziomie. W przypadku bowiem „dużych” macierzy korelacji, a tym samym licznej rodziny wnioskowań, gwałtownie maleje moc indywidualnych testów w przypadku zastosowania procedur kontrolujących FWE, a tym samym w wyniku zastosowania tych procedur otrzymujemy zbyt mało istotnych wyników. Modyfikacja metody analizy grafów, oparta na uniwersalnej procedurze Yekutieli-Benjaminiego pozwala uniknąć zadowalającej postawy procedur kontrolujących FWE. Wybór przez badacza tej miary błędu I rodzaju dla rodziny wnioskowań oznacza, że dopuszcza on i akceptuje pewien niewielki procent $\alpha 100\%$ — błędnych odrzuceń wśród wszystkich odrzuceń, a w zamian za to oczekuje mniej ostrożnej postawy niż w przypadku kontroli FWE i w efekcie większej liczby wykrytych istotnych statystycznie zależności. Z pewnością jest to lepsze rozwiązanie niż ignorowanie faktu testowania wielokrotnego.

6. TESTOWANIE WIELOKROTNE PRZY WERYFIKACJI OCEN PARAMETRÓW STRUKTURALNYCH LINIOWEGO MODELU EKONOMETRYCZNEGO

Procedury testowań wielokrotnych oparte na prawdopodobieństwach testowych można polecić również do kontroli efektu testowania jednoczesnego przy weryfikacji ocen parametrów strukturalnych liniowego modelu ekonometrycznego.

Badania empiryczne prowadzone przez autorkę (Denkowska, 2007b), dotyczące płac w okresie transformacji systemowej w Polsce, nawiązywały do badań prowadzonych przez zespół kierowany przez S. M. Kota⁴ (1999). Pokazują one w jaki sposób należy przeprowadzać kontrolę testowania wielokrotnego istotności parametrów strukturalnych w modelach regresji wielorakiej oraz jakie są skutki braku takiej kontroli.

Należy zatem mieć świadomość, że niekontrolowane testowanie równoległe wielu hipotez prowadzić może do „bardzo kiepskich modeli”. Tam, gdzie to możliwe, konieczne jest ograniczanie liczby testowań. Zaleca się również stosowanie metod wyboru zmiennych objaśniających, które nie opierają się na wnioskowaniu wielokrotnym (np. metoda Z. Hellwiga) lub kontrolują ten efekt (zmodyfikowana metoda analizy grafów, Denkowska 2007a). W przypadku bardzo bogatych zbiorów zmiennych kuszące jest wykorzystanie np. sekwencyjnych metod doboru zmiennych. Jednak na końcowym etapie budowy modelu regresji konieczne jest skorygowanie wyników poprzez kontrolę efektu testowania wielokrotnego współczynników regresji cząstkowej. W przykładzie zaprezentowanym przez autorkę (Denkowska, 2007b), metody testowań wielokrotnych zastosowane do modelu regresji, uznanego za ostateczny przez autorów *Analizy ekonometrycznej kształtowania się płac w Polsce w okresie transformacji*, pozwoliły stwierdzić, że aż trzy zmienne niesłusznie zostały uznane za istotne statystycznie. Tym samym budowa modelu regresji powinna być kontynuowana. Rezultat ten otrzymano na podstawie analizy ostatecznego modelu regresji, bez ingerowania w proces budowy modelu. Przykład ten wskazuje na konieczność kontroli efektu testowania wielokrotnego istotności współczynników regresji cząstkowej, przynajmniej na końcowym etapie budowy modelu regresji. Kontrola ta pozwoli sprawdzić, czy pozostawione parametry strukturalne nie zostały uznane za istotne wyłącznie z powodu testowania wielokrotnego podczas weryfikacji ich istotności statystycznej.

Wszędzie tam, gdzie równoległe testujemy hipotezy dotyczące parametrów regresji należy zatem stosować metody testowań wielokrotnych. Wybie-

⁴ Autorzy *Analizy ekonometrycznej kształtowania się płac w Polsce w okresie transformacji* badali metodami ekonometrycznymi mechanizmy kształtujące rozkłady płac w okresie przechodzenia od odgórnie sterowanej gospodarki do normalnie funkcjonującego rynku pracy. Wszelkie badania obejmowały między innymi estymację parametrów regresyjnych modeli płac.

rać można zarówno spośród klasycznych metod opartych na wielokrotnych przedziałach ufności (np. przedziały Scheffégo, przedziały Bonferroniego), jak również spośród rozwiązań nieklasycznych, a w szczególności z metod testowań wielokrotnych opartych na uporządkowanych prawdopodobieństwach testowych. Spośród metod testowań wielokrotnych opartych na prawdopodobieństwach testowych autorka (Denkowska, 2007b) zaleca stosowanie metod uniwersalnych: Holma i Yekutieliiego-Benjaminiego. W przypadku kontroli FWE autorka poleca procedurę Holma, natomiast gdy rozważane są bardzo bogate zbiory, kilkudziesięciu lub kilkuset zmiennych, warto spojrzeć na testowanie mniej tradycyjnie i dopuścić pewien niewielki odsetek błędnych odrzuceń w zbiorze wszystkich odrzuceń. Zastosowanie procedur kontrolujących FWE w przypadku tak bogatych zbiorów powoduje, iż indywidualne decyzje są podejmowane przy tak małych indywidualnych poziomach istotności, że bardzo rzadko dochodzi do stwierdzenia istotności parametru regresji, a wniosek końcowy wysnuty będzie na poziomie istotności α dla całej rodziny wnioskowań. W przypadku kontroli FDR, metoda Benjaminiego-Yekutieliiego jest metodą uniwersalną, którą zastosować można zawsze bez względu na zależności między statystykami testowymi. W sytuacji, gdy badania dotyczą istotności parametrów strukturalnych ortogonalnego modelu liniowego lub gdy statystyki testowe są dodatnio regresyjnie zależne (Lehmann, 1966), należy zastosować⁵ procedurę o większej mocy — procedurę Hochberga-Benjaminiego. Przy interpretacji wyników pamiętać jednak należy, że w tym przypadku α oznacza akceptowany odsetek błędnie uznanych za istotne współczynników regresji wśród wszystkich uznanych za istotne parametrów strukturalnych regresji wielorakiej.

7. PODSUMOWANIE

Na zakończenie warto wspomnieć o kierunkach badawczych zmierzających do poprawy mocy znanych procedur wnioskowań wielokrotnych, jak i wynalezienia nowych, lepszych rozwiązań. Istotną poprawę mocy procedur testowań wielokrotnych opartych na prawdopodobieństwach testowych można uzyskać poprzez zastosowanie repróbkiowania (ang. *resampling*) do oszacowania prawdopodobieństw testowych. Termin repróbkiowanie oznacza, że zaobserwowane dane są wielokrotnie wykorzystywane w analizie symulacyjnej w celu podjęcia wniosku. Prawdopodobieństwa uzyskane za pomocą technik repróbkiowania są zazwyczaj mniejsze od prawdopodobieństw testowych otrzymanych w wyniku zastosowania tradycyjnego testu *t*-Studenta. Testowanie wielokrotne oparte na repróbkiowaniu (np. Westfall i Young, 1993) wykorzystuje

⁵ Szerzej na ten temat u Benjaminiego i Yekutieliiego (2001).

empiryczną strukturę danych, a korzyść ze stosowania tej techniki jest szczególnie widoczna w przypadku silnych wewnętrznych korelacji w zbiorze danych. Wykorzystanie replikowania w testowaniu wielokrotnym propagują na przykład Westfall i Young (1993) oraz Bejamini i wsp. (2005). Metody te nie są jednak powszechnie dostępne w pakietach statystycznych, co utrudnia ich praktyczne wykorzystanie.

Kontrola FDR stwarza możliwość kontroli efektu testowań wielokrotnych w tych dziedzinach, w których typowe badania empiryczne polegają na rozpatrywaniu bardzo licznych rodzin wnioskowań. Znaczny konserwatyzm uniwersalnej procedury Yekutieli-Benjaminiego mobilizuje do bardziej wnikliwych badań nad procedurą LSU. Rozważane są różne modyfikacje poprawiające moc tej metody. Prowadzone są badania symulacyjne nad wymaganiami modelowymi procedury LSU, określającymi w jakich sytuacjach badawczych zapewni ona kontrolę FDR. W ostatnim czasie jedną z często rozważanych metod poprawy mocy procedury LSU jest metoda polegająca na wprowadzeniu dodatkowego etapu, którego zadaniem jest oszacowanie nieznanej liczby prawdziwych hipotez zerowych. W tym celu proponowane jest wykorzystanie metod graficznych, takich jak nieformalna metoda zaproponowana przez T. Schwedera i E. Spjøtvolla (1982), czy też nawiązująca do niej procedura Y. Hochberga, Y. Benjaminiego (zob. np. Westfall i wsp., 1999).

Podsumowując, należy stwierdzić, że w badaniach ekonomicznych przydatne mogą być zarówno procedury kontrolujące FWE, jak i FDR. Decyzja o odnośnię tego, która z tych miar powinna być kontrolowana, należy do prowadzącego badania i jest zależna od wielu czynników. Przede wszystkim należy określić, czy w badaniach dopuszczany jest pewien niewielki procent błędnych odrzuceń. Jeśli tak, to w zamian zyskujemy zazwyczaj więcej istotnych statystycznie wyników. Przy interpretacji uzyskanych tą metodą rezultatów należy jednak pamiętać o odsetku wyników błędnie uznanych za istotne. Z kolei kontrola FWE powoduje, że wykrywana jest często zbyt mała część istotnych statystycznie zależności. A ponieważ dalszym badaniom zazwyczaj poddawane są jedynie istotne wyniki, przy jednoczesnym zignorowaniu rezultatów, przy których stwierdzono brak podstaw do odrzucenia, to w efekcie przy zastosowaniu procedur FWE utracić możemy ważne relacje. Zastosowanie procedur kontrolujących FDR pozwala uniknąć zachowawczej postawy FWE przy zapewnieniu kontroli frakcji błędnych odrzuceń wśród wszystkich odrzuceń hipotez zerowych na zadanym poziomie. I z pewnością jest to lepszym rozwiązaniem niż ignorowanie efektu wielokrotności testowania.

Warto na koniec podkreślić, że ignorowanie efektu wielokrotności prowadzi do wykrywania wielu zupełnie przypadkowych zależności. Zależności takie są później często eksponowane w naukowych, jak również popularnonaukowych publikacjach jako ciekawe czy wręcz zdumiewające wyniki badań. To z kolei budzi sceptycyzm wobec metod statystycznych, podczas gdy źródłem nieporo-

zumienia są niewłaściwie przeprowadzone badania, nieuwzględniające efektu testowania wielokrotnego.

*I believe it [FDR] is one of the cornerstones in the bridge that „multiple comparisons” can offer between traditional statistical thinking and modern problems in Data Mining and Bioinformatics. It is crucial for us to strengthen this bridge and widen it. We should look for more such bridges, if we want Statistics remain a vital and important player in the arena of information sciences.*⁶

BIBLIOGRAFIA

- Bartosiewicz S. (red.) 1980. *Metody ekonometryczne. Przykłady i zadania*, PWE, Warszawa, 1980.
- Benjamini Y., Hochberg Y. 1995. *Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing*, Journal of the Royal Statistical Society, Ser. B 57, 1.
- Benjamini Y. 2001. *False Discovery Rate in Large Multiplicity Problems*; www.math.tau.ac.il/~ybenja/Temple.ppt
- Benjamini Y., Yekutieli D. 2001. *The Control of the False Discovery Rate in Multiple Testing Under Dependency*, Annals of Statistics 29.
- Benjamini Y., Kenigsberg E., Reiner A., Yekutieli D. 2005. *FDR Adjustments of Microarray Experiments*, www.math.tau.ac.il/~ybenja/Software/fdrame.pdf.
- Denkowska S. 1999. *Multiple Comparisons in Socio-Economic Research*, Kiev National University of Economics, Economic and Social Statistics in Transition, Proceedings of the Ukrainian, Polish and Slovak Scientific Seminar (Kiev, 20–22 October, 1998), Kiev.
- Denkowska S. 2005. *Zastosowanie procedur testowań wielokrotnych opartych na uporządkowanych prawdopodobieństwach testowych do wydzielania jednorodnych podgrup wartości przeciętnych*, Przegląd Statystyczny 52, 1.
- Denkowska S. 2006a. *Multiple Testing in a Correlation Matrix*, [w:] *A Comparative Analysis of the Socio-Economic Consequences of Transition Process in Central and Eastern European Countries*, J. Pocięcha (red.), Cracow University of Economics.
- Denkowska S. 2006b. *Monte Carlo analysis of the effectiveness of Multiple Comparison Procedures*, 13th Slovak-Polish-Ukrainian Scientific Seminar in Sväté Jur, 7–10, November 2006.
- Denkowska S. 2007a. *Modyfikacja metody analizy grafów*, [w:] *Klasyfikacja i analiza danych — teoria i zastosowania*, Prace Naukowe AE 1169, Taksonomia 14, K. Jajuga, M. Walesiak (red.), Wydawnictwo AE, Wrocław.
- Denkowska S. 2007b. *Testowanie wielokrotne przy weryfikacji ocen parametrów strukturalnych liniowego modelu ekonometrycznego*, złożony do druku.
- Denuit M., Scaillet O. 2004. *Nonparametric tests for positive quadrant dependence*, Journal of Financial Econometrics, 2004, 2, 422–450.
- Domarński Cz. 1990. *Testy statystyczne*, PWN, Warszawa.
- Domarński Cz., Pruska K. 2000. *Nieklasyczne metody statystyczne*, PWE, Warszawa.
- Grabiński T., Wydymus S., Zeliaś A. 1982. *Metody doboru zmiennych w modelach ekonometrycznych*, PWN, Warszawa.
- Goryl A., Jędrzejczak Z., Kukuła K., Osiewalski J., Walkosz A. 2003. *Wprowadzenie do ekonometrii w przykładach i zadaniach*, PWN, Warszawa.
- Hochberg Y., Tamhane A.C. 1987. *Multiple Comparison Procedures*, John Wiley & Sons, NY, 1987.

⁶ Patrz: Y. Benjamini (2001).

- Holland B., Copenhaver M.D. 1987. *An Improved Sequentially Rejective Bonferroni Test Procedure*, Biometrics 43, 417–423.
- Holm S. 1979. *A Simple Sequentially Rejective Test Procedure*, Scandinavian Journal of Statistics, 6.
- Kot S.M. (red.) 1999. *Analiza ekonometryczna kształtowania się płac w okresie transformacji*, PWN, Warszawa–Kraków.
- Lehmann E. 1966. *Some Concepts of Dependence*, Annals of Mathematical Statistics, 37, 1137–1153.
- Miller R.G. 1981. *Simultaneous Statistical Inference*, Springer-Verlag, New York.
- Schweder T., Spjøtvoll E. 1982. *Plots of p-Values to Evaluate Many Tests Simultaneously*, Biometrika 69, 3, 493–502.
- Shaffer J.P. 1995. *Multiple Hypothesis Testing*, Annual Review of Psychology 46, 561–584.
- Sokołowski A., Denkowska S. 2000. *Testowanie wielokrotne w macierzy korelacji — eksperyment symulacyjny*, [w:] *Materiały z XXXV Konferencji Statystyków, Ekonometryków i Matematyków Akademii Ekonomicznych Polski Południowej (Osieczany, 23–25 III 1999 r.)*, A. Zeliaś (red.), Kraków.
- Westfall P.H., Young S.S. 1993. *Resampling Based Multiple Testing*, Wiley, New York.
- Westfall P.H., Tobias R.D., Rom D., Wolfinger R.D., Hochberg Y. 1999. *Multiple Comparisons and Multiple Tests, Using the SAS System*, SAS Institute Inc.